BarraCNE6 model under ESG rating factor optimization based on multi-source text analysis

Group Member： FENG GUANLEI、GUO TIANKAI、JIANG XIACHENG、LI YUXIAO LIU MAORONG、LIU XINRUI、ZHANG DINGZHOU、ZHAO YICHONG、ZHOU ZIYOU、HE JUNDONG

1．Barra basis risk model based on the CNE6 model

1.1. Introduction to the model and introduction to the factors

Multi-factor risk modeling assumes that stock returns are driven by a set of common factors, so that the return forecasts of stock portfolios can be transformed into the return forecasts of factors. Specifically, with reference to the Barra CNE6 model, which introduces three main types of factors to portray stock returns, namely country, industry and style factors, the expected return of any stock can be expressed as a linear combination of the following factor exposures and factor returns.

$$R_n = f_c + \sum_{i=1}^{P} X_n^{I_i} f_{I_i} + \sum_{j=1}^{Q} X_n^{S_j} f_{S_j} + \varepsilon_n$$

Where R_n represents the expected stock return, f_c represents the expected return of the country factor. $X_n^{I_i}$ is the exposure of individual stocks on the industry, the individual stocks are exposed to 1 on the industry they belong to, and the rest of the industry is all 0. $X_n^{S_j}$ is the exposure of individual stocks on the style factor. When solving the model, the factor exposure in the above equation is the factor value calculated based on the financial data, trading data, etc. Then the stock return is taken as the current value and the factor exposure is taken as the value of the previous period for the regression solution, and the regression coefficient obtained is the factor return.

1）Country Factors

In order to strip the returns of the market as a whole from the sector and style factors, we explicitly set up country factors in the model. The concept of country factor is analogous to the intercept term in a regression model, i.e., the portion shared by different stocks that cannot be explained by industry and style. In Barra's China equity risk models CNE5 and CNE6, country factors are explicitly set, which can effectively improve the explanatory power of the models. Barra's research team pointed out in the USE4 model paper that country factors can be viewed as market portfolios weighted according to market capitalization, which can separate the market effect from the pure sector and style factors, and thus can explain the sector and style factors

more directly. style factors.

The calculation of the country factor is very simple, by setting the country factor exposure to 1 for all stocks.

2）Industry Factor

Industry is an important factor in explaining stock returns, and this paper uses CITIC Tier 1 industries as the basis for industry factor classification. The industry factors are calculated using the one-hot coding form, i.e., for the industry factor exposure of a particular stock, the industry to which the stock belongs is 1, and other industries are 0. Since the CITIC Industry Classification Standard was adjusted in December 2019 with the addition of the new Comprehensive Financials Tier 1 industry, 29 industry factors were used in the calculation of the historical data prior to the adjustment, and 30 industry factors after the adjustment.

3）Style Factors

Compared to the previous generation CNE5 model, the Barra CNE6 model, the next generation model, modifies, integrates and refines the style factor system. Specifically, the model offers two versions of the style factor system, with the Long Term Model focusing on longer-term risk management and containing 16 long-term-oriented style factors, while the Trading Model offers a more high-frequency management approach with the introduction of four additional short-term style factors Analyst Sentiment, Industry Momentum, Seasonality and Short-Term Reversal.In this paper, all 20 style factors are used to construct a multifactor risk model for the sake of comprehensiveness. In this case, the factor exposure of each style factor is synthesized from one or several descriptor variables calculated from financial data, trading data, etc., as shown in the table below. For the case of multiple descriptive variables, this paper uses an equal weighting approach to synthesize the style factors. The data sources used in this paper are WIND database and Sunrise Forever database.

Style factor	**Describing variables**	**Describe**
Analyst Sentiment	RR	Rating adjustment ratio
	ETOPF	Analysts predict changes in EP ratio
	EPSF	Analysts predict EPS changes
Beta	HBETA	Historical beta
Book-to-Price	two	book-to-market
Dividend Yield	DTOP	dividend yield
	DTOPF	Analysts predict dividend yield
Earnings Quality	ABS	Accrued items in the balance sheet
	ACF	Accrued item s in cash flow state ment

Earnings Variability	VSAL	Operating income volatility
	VERN	Profit volatility
	VFLO	Cash flow volatility
	ETOPF STD	Analysts forecast EP over standard deviation
Eanings Yield	CETOP	Cash profit to market ratio
	ETOP	EP ratio
	EM	Enterprise multiplier
	ETOPF	Analysts predict EP ratio
Growth	EGRLF	Analysts forecast long-term earnings growth rates
	EGRO	Earnings per share growth rate
	SGRO	Operating income per share growth rate
Industry Momentum	INDMOM	Sectoral Momentum
Investment Quality	AGRO	Total Assets Growth Rate
	IGRO	Growth rate of stock issuance volume
	CXGR	Capital expenditure growth rate
Leverage	MLEV	Market leverage ratio
	BLEV	Book leverage ratio
	DTOA	Asset liability ratio
Liquidity	STOM	Monthly turnover rate
	STOQ	Quarterly turnover rate
	STOA	Annual turnover rate
	ATVR	Annualized trading volume ratio
Long-Term Reversal	LTRSTR	Long term relative strength
	LTHALPHA	Long term History Alpha
Mid Ca pitalization	MIDCAP	Medium market value

Momentum	RSTR	relative intensity
	HALPHA	Historical Alpha
Profitability	ATO	Asset turnover
	GP	Gross profit margin of assets

1.2. Factorization

The basic structure and calculation methods of the three types of factors are described above, in which the country factor and industry factor exposure are both 0-1 variables. And the style factor calculation also involves data processing (taking extreme values, standardization, and missing value filling), which are described below in turn.

1）Depolarization

In order to avoid extreme values from interfering with the regression results, depolarization of the descriptive variables is generally performed. In this paper, the most common MAD absolute median deviation method is used, where the upper and lower limits are taken as 5 times the median absolute deviation, and the extreme values beyond the upper and lower limits are truncated. It should be noted that this step is skipped for some descriptive variables that are not suitable for de-extremeization treatment due to large differences in values between different stocks or a large number of missing values. Among them, three descriptor variables, IGRO stock issue growth rate, DTOP dividend yield and DTOPF analyst forecast dividend yield, are not subject to depolarization treatment.

2）Standardization

Since different descriptive variables have different computational logic and differences in magnitude and cannot be directly compared numerically, the factor values will be subjected to standardization operations. Specifically, let the original value of the descriptor variable k corresponding to stock n be X_{nk}^{Raw}, then the standardization formula is：

$$X_{nk} = \frac{X_{nk}^{Raw} - \mu_k}{\sigma_k} \qquad \mu_k = \sum_{n=1}^{N} w_n X_{nk}^{Raw}$$

Where σ_k represents the equal-weighted standard deviation and μ_k is the mean calculated based on the free float market capitalization weighting. The use of free-float market capitalization weighting, rather than the commonly used equal weighting, is in the hope that the exposure of the market benchmark portfolio to the individual style factors is 0. In addition to the descriptive variables, standardization is also performed once at each weighting of the synthetic style factors and broad category factors, thus ensuring the 0 nature of the market capitalization-weighted means

The advantage of using free float market capitalization weighting for the calculation of the mean lies in the fact that for a market benchmark portfolio (a portfolio derived from all stocks

based on free float market capitalization), the exposure to either style factor is 0, i.e.:

$$X_{Pk} = \sum_{n=1}^{N} w_n \cdot X_{nk} = \sum_{n=1}^{N} w_n \frac{X_{nk}^{Raw} - \mu_k}{\sigma_k} = \frac{1}{\sigma_k}\left(\sum_{n=1}^{N} w_n X_{nk}^{Raw} - \mu_k \sum_{n=1}^{N} w_n\right)$$

$$= \frac{1}{\sigma_k}\left(\sum_{n=1}^{N} w_n X_{nk}^{Raw} - \mu_k\right) = 0$$

3）Fill in missing values

For cases where there are missing values for the descriptive variables, the industry median is used to fill in.

4）Orthogonalization

In order to avoid the impact of covariance between the factors on the regression effect, some of the style factors were orthogonalized by referring to the original Barra CNE6 article. Among them, the Residual Volatility factor is orthogonalized to the Beta and Size factors, and the Long-Term Reversal factor is orthogonalized to the Momentum factor, respectively. In addition, since Long Term Model and Trading Model adopt different factor systems respectively, in order to ensure the consistency of regression results when using different models, it is also necessary to orthogonalize the four additional short-term factors introduced by Trading Model Analyst Sentiment, Industry Momentum, Seasonality, and Short-Term Reversal to orthogonalize the remaining 16 long-term factors.

After the treatment process of depolarization, standardization, and filling in missing values for the descriptive variables, then according to the previously described sequential weighting, synthesized to obtain the style factors; orthogonalization of the style factors, and then weighted to obtain the corresponding broad categories of factors for the subsequent model solving and factor testing.

2. Adding ESG factors to the base model

ESG (Environmental, Social and Corporate Governance) investment has become an important trend in global financial markets. The purpose of this study is to analyze the impact of ESG rating data on stock future returns (next_Rtn) and to test whether ESG factors provide additional explanatory power beyond traditional financial factors.

2.1. Data sources

ESG Rating Data Acquisition and Calculation

The explanatory variables in this paper are CSI ESG rating indicators, WIND ESG rating indicators, Business Gateway Green ESG rating indicators, Allied Wave FIN-ESG rating indicators, Bloomberg ESG rating scores, and FTSE Russell ESG rating scores, and the standard deviation of the ratings is calculated to measure the company's ESG scores and the divergence of ESG ratings.

Its CSI ESG ratings, WIND ESG ratings and Allied Wave FIN-ESG ratings are all divided into 9 grades, from low to high, as C, CC, CCC, B, BB, BBB, A, AA, AAA, and based on the above assignment method, the 9 grades of ratings C to AAA are assigned to be 1-9 in order, i.e., when the rating is C, ESG=1, and when the rating is C, ESG=1, and when the rating is C, ESG=1. When the rating is C, ESG=1, when the rating is CC, ESG=2, when the rating is CCC, ESG=3, and so on. The ESG rating of Shangdao Ronglv is divided into 10 grades, with D, C-, C, C+, B-, B, B+, A-, A, A+ from the lowest to the highest. Based on the above method of assigning values, the 10 grades of D-A+ are assigned as 0 to 9, i.e., when the grade is D, ESG=1, ESG=2, CCC, ESG=3, and so on. The 10 grades of D-A+ are assigned to 0-9, i.e., ESG=0 for a D rating, ESG=1 for a C- rating, ESG=2 for a C rating, and so on. Bloomberg ESG Ratings rounds specific scores to the nearest 10%. The FTSE Russell ESG ratings take the specific score as 200% of the sample data.

In the actual sample data, this method of assigning scores results in a similar range of values for the six ESG ratings, thus ensuring that their impact on the divergence of ESG ratings is equally weighted. After the above organization of the 6 types of ESG rating methods, this paper calculates the mean value of the ESG rating scores of the 6 types of indicators, thus obtaining the ESG score ESG_rank data. The standard deviation of the ESG rating scores of the 6 types of indicators is calculated to obtain the ESG rating divergence ESG_uncertainty data.

2.2 Calculation of existing ratings

Text Data Collection and Cleaning

Data sources include corporate disclosure documents (ESG reports, annual reports), news media (e.g., Reuters, Caixin), social media (e.g., Snowball, Twitter), and government regulatory information.

Text pre-processing includes denoising (e.g., removing advertisements and irrelevant content), standardizing terminology (e.g., unifying the expressions of “carbon emissions” and “greenhouse gas emissions”), and sentiment analysis (determining the positive/negative tendency of the text towards ESG).

Feature extraction and factorization

Keyword extraction: TF-IDF, BERT and other models are used to identify ESG-related keywords (e.g. “carbon neutral”, “board diversity”).

Topic Modeling: Use LDA (Latent Dirichlet Allocation) or NMF (Non-Negative Matrix Factorization) to extract ESG topic distributions (e.g., environmental risks, labor rights, etc.).

Sentiment and Controversy Analysis: Measure the frequency and severity of ESG-related controversial events (e.g., “environmental fines,” “excessive executive compensation”) in the text.

Multi-source data fusion and rating generation

ESG features from different data sources are weighted (e.g., 30% for corporate self-assessment, 20% for news and public opinion, and 50% for third-party data) and downscaled by Principal Component Analysis (PCA) or Factor Analysis to generate a comprehensive ESG score.

2.3 ESG factor calculation: standardization and divergence metrics

Data standardization

Due to the different rating systems of the institutions, they need to be standardized and converted to comparable values:

- CSI, WIND, Allied Wave FIN-ESG (9 grades): C=1, CC=2, ... , AAA=9.
- Merchant Road Fusion Green (10-step): D=0, C-=1, ... , A+=9.
- Bloomberg ESG: Raw scores (0-100) rounded up by 10% (e.g. 85 → 8.5 → 9).
- FTSE Russell ESG: raw score (0-5) multiplied by 200% converted to 0-10 (e.g. 3.5 → 7).

ESG composite score (ESG_rank) and divergence (ESG_uncertainty)

- ESG_rank: the average of the 6 categories of institutional scores, reflecting the overall ESG performance of companies.
- ESG_uncertainty: the standard deviation of the 6 categories of institutional scores, measuring rating divergence (the larger the value, the greater the difference in ESG ratings between institutions).

2.4 Follow-up processing: lag adjustment for ESG data

Lagged assumptions

The impact of ESG performance on the market usually has a delayed effect, so ESG data lagged by one period is used (year t data is used to explain stock returns in year t+1).

Treatment

- **Data matching:** correlating 2020 CSI ESG ratings with 2021 individual stock monthly returns, and so on.
- **Economic significance:**
  - It takes time for corporate ESG improvements to be recognized by the market (e.g., financial benefits of environmental investments may lag by 1-2 years).
  - Endogeneity issues with contemporaneous data (e.g., stock price volatility inversely affecting ESG ratings) can be avoided.

3. Final model

3.1 Base model and model after adding ESG factors

An important assumption of the classical linear regression model is that the random error terms in the regression function are homoskedastic, i.e. they all have the same variance, this is to ensure that the regression parameter estimates have good statistical properties. Although the multifactor model assumes that the idiosyncratic returns of each stock are uncorrelated with each other, the variances of the idiosyncratic returns are not the same, i.e., there is heteroskedasticity, for which the regression equation needs to be solved using the weighted least squares method. For this type of problem, the ideal regression weight is the inverse of the variance of the idiosyncratic returns, which is unknown at this point; however, based on historical experience, the variance of idiosyncratic returns is usually inversely proportional to the stock's free-float market capitalization, and therefore the square root of the free-float market

capitalization is used as the weight in the regression.

$$R_n = f_c + \sum_{i=1}^{P} X_n^{I_i} f_{I_i} + \sum_{j=1}^{Q} X_n^{S_j} f_{S_j} + \varepsilon_n$$

$$w = \left(\frac{\sqrt{s_1}}{\sum_{i=1}^{N}\sqrt{s_i}}, \ldots, \frac{\sqrt{s_N}}{\sum_{i=1}^{N}\sqrt{s_i}}\right)$$

Where s_i represents the free float market capitalization of the stock i, the number of stocks is denoted as N and the number of factors is denoted as K. In addition, since the regression model in this paper explicitly includes the country factor, the exposure of any stock in which is a constant 1, in conjunction with the fact that the sum of the exposures of a single stock to all the industry factors is obviously also 1, which brings about full covariance and results in the regression model being unsolvable. In order to solve the problem of complete covariance, it is necessary to additionally introduce a new constraint, which requires that the mean value of the industry factor weighted by the free float market capitalization is 0, s_{I_i} in the following equation represents the sum of the free float market capitalization of industry I_i.

$$\sum_i s_{I_i} f_{I_i} = 0$$

Now that we have obtained a weighted regression model with constraints, we next derive the form of the analytical solution of the model using weighted least squares with constraints, borrowing from Ishikawa et al. (2020) [6]. First, some matrix notation is introduced, R for the stock return vector; X for the factor exposure matrix with dimension $N \times K$; and W for the regression weight matrix, i.e., $W = diag(w) = diag\left(\frac{\sqrt{s_1}}{\sum_{i=1}^{N}\sqrt{s_i}}, \ldots, \frac{\sqrt{s_N}}{\sum_{i=1}^{N}\sqrt{s_i}}\right)$. The difficulty in the solution lies in how to express the constraint that the mean of the weighted industry factors is 0. For this reason, an additional constraint matrix C with dimension $K \times K-1$ needs to be constructed.The constraints are re-expressed in the following matrix constraint form.

$$\begin{bmatrix} f_c \\ f_{I_i} \\ \vdots \\ f_{I_P} \\ f_{S_1} \\ \vdots \\ f_{S_Q} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & -\frac{s_{I_1}}{s_{I_P}} & -\frac{s_{I_2}}{s_{I_P}} & \cdots & -\frac{s_{I_{P-1}}}{s_{I_P}} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} f_c \\ f_{I_i} \\ \vdots \\ f_{I_{P-1}} \\ f_{S_1} \\ \vdots \\ f_{S_Q} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

The matrix on the right-hand side of the equation above, i.e., the constraint matrix C, transforms the original constraints formally and rewrites them in order to express the returns of the last industry factor f_{I_P} in terms of a linear combination of the returns of the other industries, thus facilitating the derivation. With the above notation, the analytical solution of the regression model can be obtained based on the weighted least squares method with constraints.

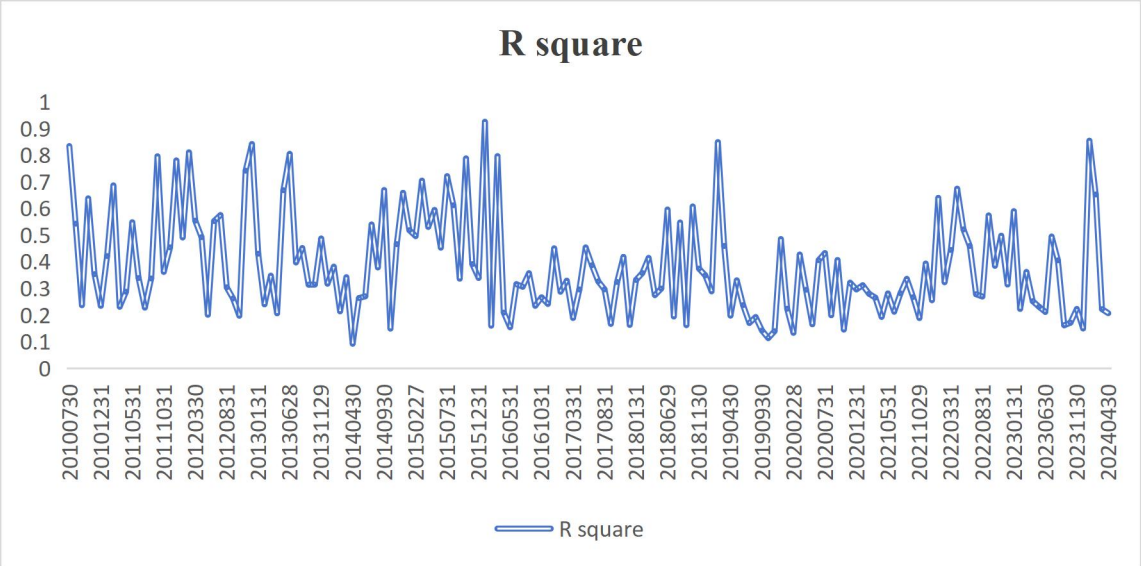
$$\hat{f} = C(C^{'}X^{'}WXC)^{-1}C^{'}X^{'}WR := W^{*}R$$

$$W^{*} = C(C^{'}X^{'}WXC)^{-1}C^{'}X^{'}W$$

Where $\hat{f}$ is the factor return estimate obtained from the regression model, which is used for subsequent return decomposition, risk prediction, etc.

Specifically, we ran monthly regressions over the period June 2010-May 2024 to obtain factor returns. The corresponding regression significance statistics and R^2 are also calculated to assess the significance level and explanatory strength of the regression models.

The regression model was solved by using the next month's returns as the dependent variable and the country factor, industry factor and style factor exposures for the last period at the end of the month as the independent variables. The following is a visualization of the results of R^2 for a total of 167 cross-sectional regressions from June 2010 to May 2024.The visualization shows that the volatility of the model regression results in terms of goodness-of-fit is high and fluctuates drastically between 0.1 and 0.9, with an average of R^2 of 0.38 for the model.



Style factor	|T|average value	percentage of |T|>2	Stabilization factor	VIF	Factor Returns Mean
Analyst Sentiment	1.77	42.17%	0.73	1.01	0.17%
Beta	2.68	63.86%	0.98	1.92	0.23%
Book-to-Price	1.43	48.19%	0.98	2.57	0.14%
Dividend Yield	1.40	32.53%	0.96	1.57	0.09%
Earnings Quality	1.19	26.51%	0.87	1.22	0.03%
Earnings Variability	0.76	19.28%	0.94	1.30	-0.03%
Earnings Yield	1.51	45.18%	0.97	2.49	0.12%

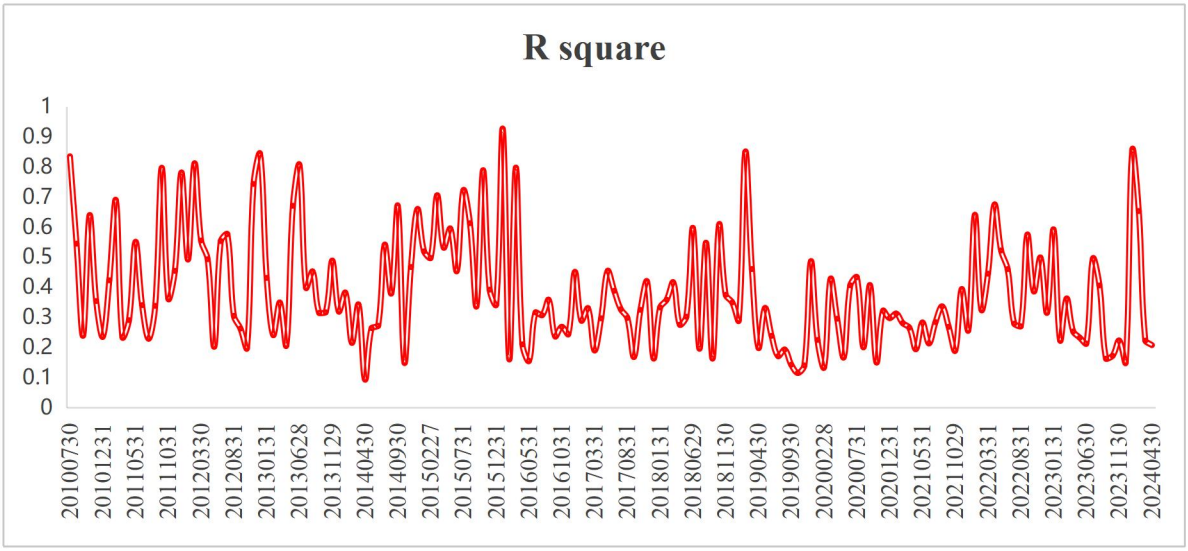
Growth	1.27	33.13%	0.96	1.49	0.07%
Industry Momentum	1.33	45.78%	0.40	1.34	-0.17%
Investment Quality	1.05	31.33%	0.98	1.61	-0.01%
Leverage	1.12	39.76%	0.99	1.45	-0.03%
Liquidity	2.22	63.25%	0.91	2.27	-0.76%
Long-Term Reversal	1.26	39.16%	0.92	1.88	0.07%
Mid Capitalization	2.46	68.67%	0.97	1.09	-0.27%
Momentum	2.62	67.47%	0.84	1.90	0.16%
Profitability	1.21	36.75%	0.98	2.28	0.04%
Residual Volatility	1.11	55.42%	0.96	2.61	-0.01%
Seasonality	1.22	33.13%	-0.07	1.01	-0.04%
Short-Term Reversal	2.17	59.04%	0.20	1.34	-0.59%
Size	2.30	75.30%	0.99	2.14	-0.33%

Significance of regression coefficients: During the test period, the |T| means of Beta, Liquidity, Mid Capitalization, Momentum, Residual Volatility, Short-Term Reversal, and Size style factors are higher than 2, indicating that these style factors explain stock returns more strongly; Earnings Quality and Earnings Variability style factors have smaller |T| means and explain stock returns less strongly. Earnings Quality and Earnings Variability have smaller |T|means, which are weaker in explaining stock returns. This also reflects the limited explanatory power of the less frequently updated value factors for stock return forecasts.

Stability coefficients: The Analyst Sentiment, Industry Momentum, Seasonality, and Short-Term Reversal style factors belonging to the Trading Model have lower stability coefficients during the test period, which is consistent with Barra's original intent of distinguishing between the Long Term Model and the Trading Model, i.e., the Long Term Model and the Trading Model have lower stability coefficients. Trading Model, which is consistent with Barra's original intent that the factor exposures of these four style factors change more rapidly and are more suitable for capturing short-term investment opportunities.

VIF: During the test period, the VIF of the style factors are all less than 5, i.e., there is no covariance.

For the results of the model after adding the ESG factors, the visualization of the goodness-of-fit is similar to the base model, with an average of 0.39 for the model, which is an improvement over the original base model.



Style factor	|T|average value	percentage of |T|>2	Stabilization factor	VIF	Factor Returns Mean
Analyst Sentiment	1.78	42.17%	0.73	1.01	0.17%
Beta	2.69	63.86%	0.98	1.92	0.24%
Book-to-Price	1.43	48.19%	0.98	2.57	0.14%
Dividend Yield	1.41	32.53%	0.96	1.57	0.08%
Earnings Quality	1.19	26.51%	0.87	1.22	0.03%
Earnings Variability	0.76	19.28%	0.94	1.30	-0.03%
Earnings Yield	1.52	45.18%	0.97	2.49	0.12%
Growth	1.28	33.13%	0.96	1.49	0.07%
Industry Momentum	1.34	45.78%	0.40	1.34	-0.17%
Investment Quality	1.06	31.33%	0.98	1.61	-0.01%
Leverage	1.13	39.76%	0.99	1.45	-0.03%
Liquidity	2.22	63.86%	0.91	2.27	-0.76%
Long-Term Reversal	1.26	39.16%	0.92	1.88	0.07%
Mid Capitalization	2.46	68.67%	0.97	1.09	-0.27%
Momentum	2.62	67.47%	0.84	1.90	0.16%
Profitability	1.21	36.75%	0.98	2.28	0.04%
Residual Volatility	1.11	55.42%	0.96	2.61	-0.01%
Seasonality	1.22	33.13%	-0.07	1.01	-0.04%
Short-Term	2.18	59.04%	0.20	1.34	-0.59%

Reversal					
Size	2.30	75.30%	0.99	2.14	-0.33%
ESG_Uncertainty	1.53	34.94%	0.95	2.26	-0.01%
ESG_Rank	1.39	28.31%	0.92	2.35	0.07%

The regression significance of the ESG factor is better than the value-based factors such as Growth, Earnings Quality, and Earnings Yield during the test period, and performs better for factors with lower frequency. It indicates that ESG factors have some ability as value-based factors at the level of stock return explanation. Meanwhile, due to the lower frequency of updating, the stability coefficient of the factors is higher and the covariance is smaller.

The improved multi-factor risk model based on ESG factors has the following applications:

- Managing portfolio style exposure. By constructing a suitable factor system, the factor exposure of each stock on each style factor can be calculated, and further the style factor exposure of the portfolio can be calculated. Thus, the portfolio style factor exposure can be brought into the portfolio optimization solution to control or adjust the specific target style exposure for active portfolio management.
- More accurate estimation of stock covariance matrix for risk prediction. Accurate estimation of the equity covariance matrix is a key issue when forecasting portfolio volatility. By transforming the high-dimensional stock covariance matrix estimation problem (4,000 stocks) into a lower-dimensional factor covariance matrix estimation problem (50 factors) through a multi-factor risk model, the accuracy of estimation can be greatly improved, thus increasing the accuracy of risk prediction.
- Ex-post imputation of the portfolio By imputing the return and risk sources of the fund products in terms of style and industry, we can gain a deeper understanding of the return and risk sources of the fund products.

3.2 Stock price prediction model based on LightGBM model

- Implementation Process

Based on the factors given by previous step, we are now focusing on the prediction of future returns with LightGBM deep learning model.

Similarly, after removing missing values and ‘Country’ factor which doesn’t provide differentiating information, we take standardized factors as explanatory variables and next month’s return as response variables. The explanatory variables include 30 bool variables as industry factors and 20 quantitative variables for other factors.

By setting observations from June 2010 to January 2024 as training dataset and observations from February 2024 to March 2024 as testing dataset, we then implement hyperparameter tuning of LightGBM model. Based on the results given by grid search, random search and Bayesian optimization, we adopt the hyperparameters below:

"learning_rate": 0.01,

"n_estimators": 50,

"num_leaves": 34,

"min_child_samples": 23,

"max_depth": 3,

"lambda_l1": 2.3,

"lambda_l2": 3.5,

"min_gain_to_split": 0,

"feature_fraction": 0.8,

"bagging_fraction": 0.8,

"bagging_freq": 5

Fitting the model with training data, we get the LightGBM model. Applying it on the testing data, we get a series of predicted returns of the next month with "MSE of test data: 0.0164, RMSE of test data: 0.1279, MAE of test data: 0.0841, R^2 of test data: 0.0020". The relatively important factors are ‘Liquidity’, ‘Mid_Capitalization’, ‘Size’, and ‘Short-term_Reversal’. LightGBM doesn’t perform well in this case.

Additionally, some of our attempts to enhance model results are proved to be ineffective. By removing the 30 industry factors, we test if having too many bool variables affects model accuracy. By removing early observations or COVID period observations, we test if observations of certain periods hide some important features of the data. By building train-test datasets within shorter periods, we test if LightGBM works better to capture short-term patterns. However, they all turned out to be insignificant.

- Interpretation of Results

Generally for LightGBM model, it works well to enhance interpretation performance of regression model. But it cannot help much if regression patterns are not significant inherently, which is exactly our case whose R2 is only 0.030 with linear regression method.

Our deduction on LightGBM result is that panel data of this scale doesn’t have some significant patterns that can be captured by this model. Maybe a single stock has its own evolution pattern, so a deep learning approach should be applied on each stock. Or maybe the general evolution pattern of the market exists within a month’s scale, so that we need data of higher frequency to build deep learning model.。

3.3 Revenue prediction model based on LSTM and GRU models

To address the possible data leakage of dividing the training set and test set for time series data, I separate the training set and test set by the time point, i.e., I use the data before December 31, 2022 as the training set, and the later as the test set, which ensures that data leakage will not

occur. After that, I set epochs=10, batch_size=32 to ensure the randomness of the extracted samples. The data in the training set is then extracted to train the model and the effect is tested in the test set.

Since the recurrent neural network class model captures the relationship of the time series, the input is set as a three dimensional tensor. In this model, the training set is (426410, 3, 50), i.e., the time step is 3 and the input features are 50, totaling 426410 samples. The test set is (72263, 3, 50). Thus this model uses 3 time-step features to predict the label at the next time point.

At the model specific construction level, for the GRU model:

Model Structure:

Use two-layer GRU network to process the time series data, together with Dropout layer for regularization, and finally output the prediction results through the fully connected layer.

Tuning method:

Keras Tuner's stochastic search is used to search for the optimal number of GRU units, Dropout rate, and learning rate in a preset hyperparameter space, with the goal of minimizing the loss of the validation set.

The parameter tuning results are:

GRU 1: 64 GRU 2: 32 Dropout 1: 0.30000000000000004 Dropout 2: 0.2 lr: 0.001

Anti-overfitting method:

With the Dropout layer, EarlyStopping callback and ReduceLROnPlateau callback, the model fights overfitting and improves generalization during training.

For the LSTM model, I used a similar approach.

The training set is trained using the optimal parameters and the effect is tested using the test set with the model specific parameters saved in the GRU_Model.h5 and LSTM_Model.h5 files. The final results are as follows:

GRU：

MSE: 0.015842

RMSE:0.125866

MAE: 0.086968

y_mean :-0.007134583690839367,y_std:0.12126015501029316

LSTM：

MSE: 0.015331

MAE: 0.08507

y_mean :-0.007134583690839367,y_std:0.1212601550102931

Although the errors in the training and validation phases appear to be low, the test results show that the model fails to effectively capture the pattern of returns in actual forecasts, suggesting weak generalization.The RMSE has a similar standard deviation to the target return, suggesting that the model's forecast error is close to the magnitude of the fluctuations in the data itself.

As for the reasons for the poor results, we believe there are several aspects:

1. Stock returns tend to be affected by various random factors in the market, with more data noise, and it may be difficult for a single model to extract an effective pattern, resulting in a larger prediction error.
2. The time series of the stock market is usually non-stationary and highly volatile, and the traditional GRU and LSTM models may need more adjustments or combination of other models (e.g., hybrid models, integrated methods) to improve the prediction effect when facing this kind of data.
3. The input features may not be comprehensive enough or not sufficiently extracted and transformed, and lack of effective capture of market dynamics, macroeconomic factors, etc., making it difficult for the model to make accurate predictions.